FIRST ORDER AMBISONICS DOMAIN SPATIAL AUGMENTATION FOR DNN-BASED DIRECTION OF ARRIVAL ESTIMATION

Luca Mazzon[1,2], *Yuma Koizumi*[1], *Masahiro Yasuda*[1], *and Noboru Harada*[1]

[1]NTT Media Intelligence Laboratories, Tokyo, Japan
[2]University of Padova, Padua, Italy

ABSTRACT

In this paper, we propose a novel data augmentation method for training neural networks for Direction of Arrival (DOA) estimation. This method focuses on expanding the representation of the DOA subspace of a dataset. Given some input data, it applies a transformation to it in order to change its DOA information and simulate new potentially unseen one. Such transformation, in general, is a combination of a rotation and a reflection. It is possible to apply such transformation due to a well-known property of First Order Ambisonics (FOA). The same transformation is applied also to the labels, in order to maintain consistency between input data and target labels. Three methods with different level of generality are proposed for applying this augmentation principle. Experiments are conducted on two different DOA networks. Results of both experiments demonstrate the effectiveness of the novel augmentation strategy by improving the DOA error by around 40%.

Index Terms— First Order Ambisonics, direction of arrival, deep learning, data augmentation

1. INTRODUCTION

Direction of arrival (DOA) estimation is the task of detecting the spatial position of a sound source with respect to a listener. The approaches that has been adopted to solve this problem can be classified in two main categories: parametric-based methods, like multiple signal classification (MUSIC) [1] and others [2–4], and deep neural network (DNN)-based methods [5–17]. DNN-based models often combine DOA estimation with other tasks such as sound activity detection (SAD), estimation of number of active sources and sound event detection (SED) [11–13]. In particular, Sound Event Localization and Detection was the task 3 of Detection and Classification of Acoustic Scenes and Events 2019 Challenge (DCASE2019 Challenge) [18].

In machine learning, *data augmentation* is an effective strategy to overcome the lack of data in the training set and prevent overfitting. For example SpecAugment [20], a recently published augmentation method based on time warping and time and frequency block masking of the spectrogram, achieved state of the art performance on the Speech recognition task. DCASE2018 Task2 Challenge (about audio tagging) winner [21] used mixup augmentation [22].

While data augmentation is effective for sound event detection and similar tasks, none of the documented strategies is capable of effectively increasing the spatial representativeness of a dataset, i.e. increasing the number of DOAs represented in the dataset. The critical point of the problem is that when the observed signals are modified by a data augmentation method, it must be guaranteed that the relationship between the DOA information carried by the signal and the corresponding labels is maintained. For example, augmentation techniques such as SpecAugment, phase-shifting and mixup can indeed influence DOA, although it's hard to analytically compute the new true DOA labels. In fact, according to the technical reports of DCASE 2019 task3, SpecAugment has affected adversely for DOA estimation even though it is effective for SED [19, 23].

In this paper, we propose *FOA Domain Spatial Augmentation*, a novel augmentation method based on the well-known rotational property of First Order Ambisonics (FOA) sound encoding. The basic idea of the method is to apply some transformations to the FOA channels (and corresponding labels) to modify and simulate a new DOA of the recorded sounds in a predictable way. Such transformations are: channel swapping and inversion, application of a rotation formula (i.e. Rodrigues' rotation formula) and multiplication by an orthonormal matrix, which correspond to rotations and reflections of the sound sources positions with respect to a reference system centered on the listener.

2. FIRST ORDER AMBISONICS

First-Order Ambisonic (FOA) is a digital audio encoding which describes a soundfield [24]. It has origin in the B-Format, which encodes the directional information on four channels W, X, Y and Z [24]. W carries omnidirectional information, while channels X, Y and Z carry the directional information of the sound field along the Cartesian axes of a reference system centered on the listener [24].

Adopting the same notation and convention of the dataset used for the following experiments [25], the spatial responses (steering vectors) of the FOA channels are $H_1(\phi, \theta, f) = 1$, $H_2(\phi, \theta, f) = \sqrt{3} * \sin\phi * \cos\theta$, $H_3(\phi, \theta, f) = \sqrt{3} * \sin\theta$, and $H_4(\phi, \theta, f) = \sqrt{3} * \cos\phi * \cos\theta$, where ϕ and θ are the azimuth and elevation angles of a sound source, f is frequency and $*$ is used for the multiplication operation. As it is noticeable from the expressions, FOA channels can be seen as the projections of the sound sources to the three dimensional Cartesian axes, with H_1 corresponding to channel W, H_2 to channel Y, H_3 to channel Z and H_4 to channel X. Thus, indicating with $\mathbf{S} = \{S_1, ..., S_n\}$ a set of sound sources in their STFT domain, FOA channels can be written as a sum of each source and its steering vector, that is $X = \frac{1}{N}\sum_{n=1}^{N} H_4(\phi_n, \theta_n, f) * S_n$, where $N = |\mathbf{S}|$, and ϕ_n and θ_n are the azimuth and elevation of S_n, respectively.

3. FOA DOMAIN SPATIAL AUGMENTATION

The goal of the method is, from the audio recordings in the dataset, to generate new ones with different DOA information. More specif-

https://doi.org/10.33682/3qgs-e216

ically, the problem consists in simulating a new set of spatial responses $\{H_i(\phi'_n, \theta'_n, f)\}_{i=2}^4$ corresponding to new DOA labels $\{\phi'_n, \theta'_n\}_{n=1}^N$ for the audio recordings by applying a transformation directly to the FOA channels. It is a known property of FOA that, since it encodes a soundfield rather than the sources themselves, it is possible to apply some operations directly to the channels [24], such as rotations and reflections. There are several ways to apply these transformations, leading to different augmentation strategies with different pros and cons. In the following, three strategies are proposed and compared.

3.1. First method: 16 patterns

The *16 patterns* method simply consists in applying to the data one of the 16 prefixed channel transformations summarized in Table 1, where ← indicates an assignment. The basic operations used in this method are channel swapping (e.g. $X' \leftarrow Y, Y' \leftarrow X$) and channel sign inversion (e.g. $Z' \leftarrow -Z$) or a combination between the two. Using this set of operations, it is possible to obtain 8 rotations about the z axis and 2 reflections with respect to the $xy : z = 0$ plane, for a total of 16 augmentation patterns (i.e. 15 new patterns plus the original one). The corresponding transformations for the labels are also reported in Table 1. In particular, the listed transformations correspond to the translations of $+0$, $+\pi$, $+\frac{\pi}{2}$ and $-\frac{\pi}{2}$ of the azimuth angles ϕ and $-\phi$ and to the pair of opposites ϕ and $-\phi$.

The main advantage of this algorithm, other than it's simplicity and straightforward implementation, is the possibility of it being applied to many pre-computed features, such as logmel magnitude spectrogram or phase spectrogram, since the corresponding transformations in the feature-domain are straightforward to compute (channel swapping maps to the same channel swapping, channel sign inversion maps to identity for magnitude and to a 180 degrees difference for phase). Another advantage is that it is easy to control that mapped angles remain in the same domain as the original ones. For example, in the dataset in use for DCASE2019 Challenge task3, all angles are multiples of 10 degrees and elevation angles range from -40 to $+40$ degrees. It is easy to see that the augmented angles maintain the same domain. One more important advantage of this method is that it can be applied independently on the number of the maximum number of overlapping sound sources, which is a complication for the next proposed method.

3.2. Second method: Labels First

In *Labels First* method, the basic idea is to first decide the target augmented labels, than to apply a transformation to the data accordingly. The critical aspect of this method is that while for azimuth this is always possible independently on the number of overlapping sources, it isn't the same for elevation. The reason is that when modifying the azimuth coordinates by a fixed amount by means of a rotation, *z-axis* is the common rotational axis for all the sources, while for modifying only the elevation coordinate by a fixed amount by means of a rotation, for each source, an appropriate rotation axis must be selected.

Keeping into consideration this critical aspect, assuming at first to have sound files with non-overlapping sound events, the proposed algorithm for this method is follows. For convenience, it is divided in two steps in which azimuth and elevation are augmented separately. In the first step, at first a random angle α is selected and used to translate, at each time step t with arbitrary range, the azimuth labels:

$\alpha \leftarrow \texttt{random}(0, 2\pi)$
$\phi'_t \leftarrow \phi_t \oplus \alpha$

where $\oplus$ here indicates an addition with a wrap-around on the domain $(-\pi, \pi)$, i.e. $(\phi_t + \alpha + \pi) \bmod 2\pi - \pi$. At this point, the rotation matrix around *z-axis* R_z is computed:

$$R_z = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

and applied to the channels at each time step t:

$$\mathbf{v}'_t = R_z \mathbf{v}_t, \tag{2}$$

where $\mathbf{v}_t = (X_t, Y_t, Z_t)^\top$ denotes original channels and $\mathbf{v'}_t = (X'_t, Y'_t, Z'_t)^\top$ denotes azimuth-augmented channels. In the second step, the elevation coordinate is augmented. At first, a random augmentation angle β is selected. To do so, elevation labels in the selected time range (e.g. a batch) is inspected and maximum and minimum values M_e and m_e are extracted. The elevation angle, by definition, has range $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, but in some datasets like the one in use for DCASE2019 Challenge task3, it might have a custom range (m_{er}, M_{er}). In order not to go out of this range, the augmentation angle β is extracted randomly in the range $(m_{er} - m_e, M_{er} - M_e)$[1]. At this point, elevation labels are updated:

$\beta \leftarrow \texttt{random}(M_{er} - M_e, m_{er} - m_e)$
$\theta'_t \leftarrow \theta_t + \beta$

Secondly, at each time step t, the rotation axis for augmenting elevation is computed. This axis is defined by the unit vector perpendicular to the one along the azimuthal axis, oriented properly so that a rotation of the audio channels by an angle β corresponds to the same increment of the elevation label. It can be easily verified with the right-hand rule that this unit vector corresponds to the azimuthal one rotated by $-\frac{\pi}{2}$ about the *z-axis*:

$$\mathbf{u}_t = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\phi'_t \\ \sin\phi'_t \\ 0 \end{pmatrix} = \begin{pmatrix} \sin\phi'_t \\ -\cos\phi'_t \\ 0 \end{pmatrix}, \tag{3}$$

where the first term means $R_z(-\pi/2)$. Now, Rodrigues' rotation formula is applied to the $\mathbf{v'}_t$, we obtain full-augmented channels:

$$\mathbf{v''}_t = \mathbf{v'}_t \cos\beta + (\mathbf{u}_t \times \mathbf{v'}_t)\sin\beta + \mathbf{u}_t\,(\mathbf{u}_t \cdot \mathbf{v'}_t)(1 - \cos\beta), \tag{4}$$

where $\times$ and $\cdot$ denote the cross-product and the inner-product, respectively.

The main advantage of this method is the high control over the augmented labels. For example, it allows for generating new labels which belong to the same domain of the original ones (e.g only multiple of 10° and elevation restricted to the range $(-40^\circ, 40^\circ)$, such as in the dataset used for the experiments [25]. The main disadvantage is that it is best suitable for non-overlapping sound events. There are some workarounds to adapt it to sound recordings with multiple overlapping sound events, though. Some possibilities are to apply it only to time frames with one event, to check for the somewhat rare cases in which all of time events share the same azimuth coordinates or to apply a hybrid strategy such as applying the 16 patterns method only for elevation augmentation or considering only one of the overlapping sources and computing labels for the others sources as in Channels First.

[1] In order to maximize the augmentation range, one could segment the audio recordings in frames containing the single sources. Alternatively, one could accept to extend the elevation domain of the dataset and agnostically select a fixed range for the augmentation angle β, which is convenient when augmenting an entire audio file altogether, as done in experiment 2.

Table 1: Sixteen patterns of simple spatial augmentation. Swap(X, Y) denotes $X' \leftarrow Y$ and $Y' \leftarrow X$.

	$\phi - \pi/2$	ϕ	$\phi + \pi/2$	$\phi + \pi$
θ	Swap$(-X, Y)$	original	Swap$(X, -Y)$	Swap$(-X, -Y)$
$-\theta$	Swap$(-X, Y)$, $Z' \leftarrow -Z$	$Z' \leftarrow -Z$	Swap$(X, -Y)$, $Z' \leftarrow -Z$	Swap$(-X, -Y)$, $Z' \leftarrow -Z$
	$-\phi - \pi/2$	$-\phi$	$-\phi + \pi/2$	$-\phi + \pi$
θ	Swap$(X, -Y)$	$Y' \leftarrow -Y$	Swap(X, Y)	$X' \leftarrow -X$
$-\theta$	Swap$(-X, -Y)$, $Z' \leftarrow -Z$	$Y' \leftarrow -Y, Z' \leftarrow -Z$	Swap(X, Y), $Z' \leftarrow -Z$	$X' \leftarrow -X, Z' \leftarrow -Z$

3.3. Third method: Channels First

Channels first is the most general case of FOA Domain Spatial Augmentation. This method doesn't depend on the number of overlapping sources, but the control over labels is almost completely lost.

The procedure is as follows. A random (3×3) orthonarmal matrix R is selected. An orthonormal matrix R is a matrix such that $HH^\top = I$ and $det(H) = \pm 1$. This can be done by selecting a random (3×3) matrix and then orthonormalizing it with the Graham-Schmidt method. Augmented channels $\mathbf{v}'$ are then computed as:

$$\mathbf{v}' = R\,\mathbf{v}. \tag{5}$$

The same transformation is also applied to the labels $y = (\phi, \theta)^\top$, in Cartesian coordinates[2]:

$y_c \leftarrow \mathrm{to\_cartesian}(y)$
$y'_c \leftarrow R\,y_c$
$y' \leftarrow \mathrm{to\_spherical}(y'_c)$

An orthonormal matrix expresses a general rotoreflection. This method allows generating the most number of augmentation patterns for any number of sources, but, since there is few to none control over the labels, it is recommended to use only with datasets without any restrictions on the labels' domain, as justified by the results of experiment 2.

4. EXPERIMENT

4.1. Experimental setup

We conducted our experiments, referred to as *Experiment 1* and *Experiment 2*, using two different DOA networks, one simpler, one more sophisticated, here referred to as *Simple DOAnet* and *Sophisticated DOAnet*. Both networks give as output a single pair of azimuth and elevation angles computed in a regression fashion and a sound activity detection value computed in a classification fashion. Both networks are trained using the maximum overlapping 1 audio files of the DCASE2019 Challenge dataset [25], and evaluated on DOA error (Er) and Frame-recall (FR). We used only overlap 1 files in order to be able to evaluate the effectiveness of FOA Domain Spatial Augmentation specifically for the DOA estimation task. In systems that are able to localize more than one overlapping source, such as SELDnet [11], other tasks, such as Sound Event Detection (SED), might be influenced by the augmentation strategy and at the same time influence the performance on DOA estimation.

4.1.1. Experiment 1

Simple DOAnet has a convolutional recurrent neural network (CRNN) as a core structure, as in [10–12, 19, 26]. Input features are logscale Mel-magnitude spectrogram (logmels) and Generalized Cross-Correlation Phase Transform (GCC-PHAT) of the mutual channels, as in [12, 26]. All wav-files were downsampled at a sampling rate of 32 kHz. The length of the short-time-Fourier-transform (STFT) and its shift length were 1024 and 640 (20 ms) points, respectively. The dimension of Mel bins for logmels and GCC-PHAT was 96. The DNN structure is a CRNN, similar to a SELDnet [11] without the SED branch and with a single class DOA output. The CRNN consists of 3 convolutional neural network (CNN) layers, 2 gated recurrent unit layers, and 2 fully-connected (FCN) layers, with the total number of parameters of 545K.

As a loss function, we compute the mean average errors (MAE) between true and predicted labels for both azimuth and elevation and mask them with the true sound activity labels, then sum them to the binary cross-entropy loss of the sound activity output. The model is trained adopting the four cross-validation folds defined in [25] for 400 epochs each and selecting the best model among the epochs according to the best validation loss. The conducted experiments on this model are 3: the first is without using FOA Domain Spatial Augmentation (No Aug), the second is applying the Labels First method on 50% of the input data (LF Half) and the third is applying the Labels First method on all of the input data (LF Full). Augmentation is applied on minibatches of 100 STFT frames (2s).

4.1.2. Experiment 2

Experiment 2 is conducted on *Sophisticated DOAnet*. *Sophisticated DOAnet* is a combination method of parametric-based and DNN-based DOA estimation [27]. Sound intensity vector (IV)-based DOA estimation is used as a base method and two CRNNs are used for denoising and dereverberation of IVs. Each CRNN consists of 5 CNN layers, 2 FCN layers, and 1 bidirectional long short-term memory layer, and the total number of parameters of *Sophisticated DOAnet* is 2.79M. The details of *Sophisticated DOAnet* are described in [27].

Training was performed on the standard cross-validation folds, and selecting the best model among the epochs according to the best validation loss. Four different runs of the training are performed on this network, one without augmentation (No Aug) and one for each of the methods described in Section 3: 16 Patterns (16P), Labels First (LF) and Channels First (ChF). Based on the results of *Experiment 1*, all data augmentation was performed on 50% of the input data directly on the full length wav files. For the Labels First method, elevation augmentation angle β was selected randomly between $-20°$ and $20°$, extending the range of elevation to $(-60°, 60°)$.

4.2. Results

Experiment 1 has mainly two purposes: demonstrate the effectiveness of FOA Domain Spatial Augmentation on training a simple

[2] Since distance from the listener is not relevant for the task, when converting to and from cartesian coordinates, we always assume the norm $r = 1$, that is we consider direction of arrivals as points on the unit sphere.

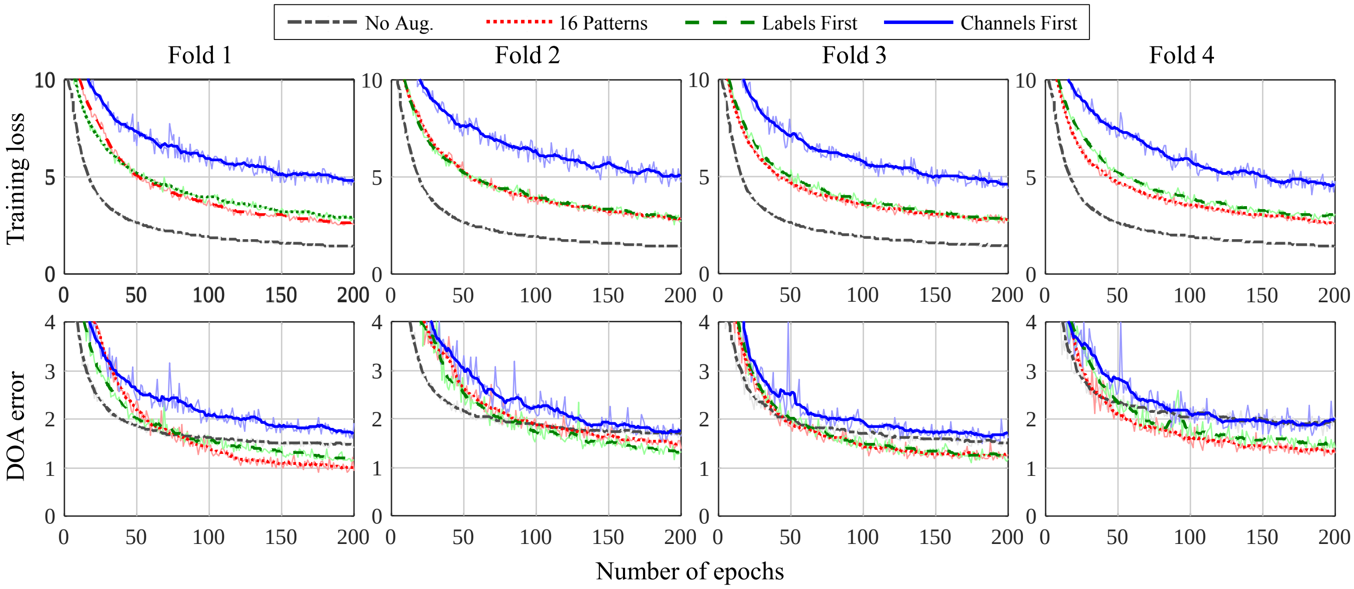

Figure 1: Training progress graph of experiment 2; training loss (top) and DOA error of validation set (bottom). It is apparent that the 16 Patterns method and the Labels First method performed better than without augmentation. The Channels First method lead to worse results, supposedly due to the over-extension of the labels domain and the consequent complication of the problem.

Table 2: Results of *experiment 1* on *Simple DOAnet*

		Fold 1	Fold 2	Fold 3	Fold 4	Ave.
No	Er	5.32	4.85	5.56	5.07	5.22
Aug.	FR(%)	97.78	98.46	97.79	97.67	97.93
LF	Er	**3.34**	**3.28**	**3.27**	**3.07**	**3.22**
Half	FR(%)	98.16	**98.89**	98.28	98.14	98.37
LF	Er	3.53	3.64	3.53	3.21	3.48
Full	FR(%)	**98.18**	98.74	**98.41**	**98.38**	**98.43**

Table 3: Results of *experiment 2* on *Sophisticated DOAnet*

		Fold 1	Fold 2	Fold 3	Fold 4	Ave.
No	Er	1.69	1.53	1.81	1.60	1.66
Aug.	FR(%)	96.91	96.46	97.14	97.50	97.00
16P	Er	**0.96**	**1.30**	**1.45**	**1.21**	**1.21**
	FR(%)	97.30	97.00	97.53	**97.70**	**97.39**
LF	Er	1.31	1.39	1.49	1.43	1.40
	FR(%)	**97.34**	94.16	**97.61**	97.14	96.56
ChF	Er	1.99	1.35	1.98	1.61	1.73
	FR(%)	96.45	**97.28**	97.24	96.96	96.98

DOA network and discover whether it is best to apply augmentation to all the data or, heuristically, to only half of the data. As reported in Table 2, both runs with the use of augmentation outperformed the run without using augmentation on all the cross-validation folds, decreasing the DOA error by 2° on average and increasing the Frame Recall of 0.5% on average. DOA error achieved the best results by augmenting 50% of the input data (0.24° better on average with respect to 100%), while augmenting all of the input data achieved the best result in terms of Frame Recall (0.06% better on average), although the results of these two runs were very close to each other.

Experiment 2 has the purpose of comparing the different FOA Domain Spatial Augmentation methods illustrated in Table 3 with each other as well as with non augmented data. The results reported in Table 3 show that in this case the 16 Patterns one was the best performing method, followed by the Labels First method, also scoring better than without augmentation in terms of DOA Error. As we expected Labels First to be the method achieving the best scores, we believe the penalty with respect to the expectation is due to the expansion of the labels domain, which means a more difficult problem to solve. As expected, the Channels First method was the least effective on this dataset, scoring worse with respect to non augmented data. It is safe to say that the determining factor for the underperformance is the too big of a difference in the labels domains of the augmented data and of the original data. In terms of frame recall, again 16 Patterns achieve the best score, although there aren't any particularly noticeable differences. In Figure 1, the training progress graphs of experiment 2 are reported. It can be clearly seen that in all the cross-validation folds there is a point since which DOA error on validation split is better with the 16 Patterns and Labels First methods rather than without augmentation.

5. CONCLUSIONS

In this paper, FOA Domain Spatial Augmentation, a novel data augmentation strategy, has been proposed. The basic idea of the method is to apply rotational transformations to the FOA channels and corresponding labels. We proposed three types of such transform: channel swapping and inversion, application of a rotation formula, and multiplication by an orthonormal matrix. It has been proven effective for training two different neural networks for the task of DOA estimation, improving the DOA error by 40%. Future research will be to further investigate the effectiveness of this augmentation strategy in different scenarios, for example with a dataset including all the possible DOAs or with overlapping sound events.

6. REFERENCES

[1] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation" in *IEEE Transactions on Antennas and Propagation*, vol. AP-34, pp. 276–280, Mar. 1986.

[2] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach" in *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, 2001.

[3] M. S. Brandstein and H. F. Silverman, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997.

[4] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 37, no. 7, 1989.

[5] M. Yiwere and E. J. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks in *International Journal of Applied Engineering Research*, vol. 12, no. 22, 2017.

[6] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[7] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment" in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

[8] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilisitic neural network" in *IEEE Transactions on Industrial Electronics*, vol. 29, no. 1, 2017.

[9] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks" in *Deutsche Jahrestagung für Akustik (DAGA)*, 2015.

[10] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network" in *European Signal Processing Conference (EUSIPCO)*, 2018.

[11] S. Adavanne, A. Politis, J. Nikuunen, and T. Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks" in *Journal of Selected Topics in Signal Processing*, 2018.

[12] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy" *arXiv preprint, arXiv:1905.00268*, 2019.

[13] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks in *Audio Engineering Society Convention 138*, 2015.

[14] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms" in *Microphone Arrays*, Springer, 2001.

[15] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[16] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models" in *Journal of Robotics and Mechatronics*, vol. 29, no. 1, 2017.

[17] W. He, P. Motlicek, and J. M. Odobez, "Deep neural networks for multiple speaker detection and localization" in *International Conference on Robotics and Automation (ICRA)*, 2018.

[18] `http://dcase.community/challenge2019/task-sound-event-localization-and-detection`.

[19] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of CRNN models" in *Tech. Report of Detection and Classification of Acoustic Scenes and Events 2019 Challenge (DCASE2019 Challenge)*, Jun. 2019

[20] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, Ekin D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition" in *Interspeech 2019*, Apr. 2019

[21] I. Jeong and H. Lim, "Audio tagging system using densely connected convolutional networks" in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, Nov. 2018

[22] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: beyond empirical risk minimization" in *International Conference on Learning Representations (ICLR)*, Apr. 2018

[23] J. Zhang, W. Ding, and L. He, "Data augmentation and prior knowledge-based regularization for sound event localization and detection," in *Tech. Report of Detection and Classification of Acoustic Scenes and Events 2019 Challenge (DCASE2019 Challenge)*, Jun. 2019

[24] F. Hollerweger, "An introduction to higher order ambisonics", `https://pdfs.semanticscholar.org/40b6/8e33d74953b9d9fe1b7cf50368db492c898c.pdf` (last access, July 17, 2019), Oct. 2008

[25] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection" in *Submitted to Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019.

[26] L. Mazzon, M. Yasuda, Y. Koizumi, and N. Harada, "Sound event localization and detection using FOA domain spatial augmentation", in *Tech. Report of Detection and Classification of Acoustic Scenes and Events 2019 Challenge (DCASE2019 Challenge)*, Jun. 2019

[27] M. Yasuda, Y. Koizumi, L. Mazzon, S. Saito, and H. Uematsu, "DOA estimation via DNN-based dnoising and dereverberation from sound intensity vector" *arXiv preprint*, 2019.